Технический отчет “Иерархическая мультимодальная тематическая модель коллекции научно-популярных текстов”

И. В. Ефимова
efimova@phystech.edu
Московский физико-технический институт, ФУПМ, Кафедра интеллектуальных систем

Рассматривается задача построения иерархической тематической модели коллекции документов. Особенность данной задачи заключается в наличии дополнительной метаинформации документов, которая включает в себя и часть тем, присутствующих в документах. В работе предложен метод послойного построения иерархии коллекции текстов. Также предложены критерии качества, оценивающие построенную модель. В экспериментах показано, что данный метод позволяет строить интепретируемые мультимодальные тематические иерархии, в которых удобно ориентироваться пользователю.

Ключевые слова: *вероятностное тематическое моделирование; аддитивная регуляризация; иерархическая модель; мультимодальная модель; научно-популярные тексты*

1 Введение

В эпоху информационных технологий появляется доступ к неограниченному объему знаний, доступных через сеть Интернет, но физически человек способен ознакомиться только с малой частью этих данных. Возникает потребность в системах автоматической организации информации для пользователя.

Последние пятнадцать лет активно развивается раздел машинного обучения, решающий задачу поиска тем в коллекции документов — вероятностное тематическое моделирование. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему. Иными словами, модель задает компактное представление для коллекции, которое позволяет быстрее ознакомиться с ее содержанием. Однако на больших текстовых коллекциях, когда число тем становится равным нескольким сотням или тысячам, даже такое представление в виде набора тем перестает быть удобным. Появляется потребность в построении иерархических тематических моделей, в которых крупные темы постепенно дробятся на более узкие, специализированные темы. Таким образом, иерархические тематические модели помогают представить структуру коллекции текстовых документов в виде иерархии тем. Это позволяет пользователю наиболее полно познакомиться с областью знаний, к которой относится коллекция.

Большинство подходов к построению иерархий вероятностные: в них термины, темы и документы считаются случайными величинами, а коллекция моделируется с помощью процесса порождения слова в документе. Одна из первых таких иерархических моделей предложена в [1]: иерархия представляется в виде дерева тем, и ее можно достраивать при добавлении новых документов в коллекцию. В [2] ключевая идея состоит в отказе от ограничения на граф: иерархия является многодольным графом, то есть темы могут иметь несколько надтем. Авторы [3] также представляют иерархию в виде многодольного графа и описывают две модели, которые автоматически определяют количество тем и количество уровней, или долей в графе. Одна модель строит иерархию документов, другая – иерархию терминов, совместить эти модели в одной не предлагается. Аналогично, в [4] темы описываются только лексикой, то есть связь документов и тем не моделируется;

ключевая особенность – темы представляются как список фраз, а не отдельных терминов, в результате повышается интерпретируемость тем. Список терминов родительской темы получается объединением списков терминов дочерних тем. Этот подход развивается в [5], где модель учитывает не только текстовую, но и иную информацию, представленную в коллекции: авторов, метки времени, локации на карте и т. д. В [6] делают акцент на трех приоритетах: масштабируемость, то есть быстрое построение модели на больших коллекциях, устойчивость, то есть построение похожих моделей при повторных запусках, и интерпретируемость. В [7] к этому списку добавляется еще одна цель: возможность учитывать указания эксперта, например указание объединить две темы. На важность масштабируемости алгоритма обучения также указывают авторы [8].

В данной работе для построения иерархических тематических моделей используется метод послойного построения иерархии, описанный в [9], который основан на аддитивной регуляризации тематических моделей (Additive Regularization of Topic Models, ARTM) [10]. В работе предложены метрики качества для оценки построенной иерархии: ошибки перого и второго рода, функционал. Эксперименты показали, что предложенный метод позволяет сторить интерпретируемые иерархии коллекций.

2 Плоская тематическая модель

2.1 Обозначения и определения

Пусть D – множество текстовых документов, W – множество всех употребляемых в коллекции терминов — слов или словосочетаний, также именуемое словарем. Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, ..., w_{n_d})$, принадлежащих словарю W. В тематическом моделировании принимается гипотеза о том, что порядок терминов в тексте не важен для определения его тематики. Тогда коллекцию можно представить в виде матрицы частот слов с элементами n_{dw} – частота вхождения термина w в документ d. Длину документа будем обозначать n_d. По матрице частот слов можно оценить вероятности появления терминов в документе: $p(w|d) = \frac{n_{dw}}{n_d}$.

Если документ, помимо текстового описания, характеризуется другими признаками, например автором или ключевыми словами, то говорят, что в него входят элементы разных модальностей. Каждой модальности соответствует отдельный словарь, состоящих из всевозможных значений элементов модальности. В примере выше словари авторов и ключевых слов будут состоять соответственно из фамилий всех авторов, написавших документы, и всех ключевых слов, встретившихся в коллекции. В этом случае считается, что термины также формируют отдельную, текстовую, модальность. Множество модальностей будем обозначать M, а их словари — $W^m, m \in M; W = \cup_{m \in M} W^m$.

Тематической иерархией мы называем многодольный граф с вершинами-темами и ребрами, показывающими связи между ними. Доли графа будем считать упорядоченными в порядке роста количества тем-вершин в каждой доле и называть их уровнями или слоями. Если две темы соединены ребром, то тему, находящуюся на уровне выше, называют родительской, а тему на следующем уровне — дочерней темой. Если самый верхний уровень иерархии представлен одной темой, ее называют корневой. Уровни иерархии будем обозначать $1, ..., L$, множества тем этих уровней — $S_1, ..., S_l$. Обычные, не иерархические, тематические модели, в которых все темы равносильны и формируют одно множество тем S, называют плоскими.

2.2 Аддитивная регуляризация тематических моделей

Модель ARTM – модификация модели PLSA, которая поддерживает многомодальность и способна учитывать одновременно несколько регулрязаторов [10].

В ARTM вводится несколько матриц тем (по одной на каждую модальность): $\Phi^m = = \{p(w|s)\}_{w\in W^m, s\in S}$.Обозначим $\Phi = \cup_{m\in M}\Phi^m$.

Также вводятся регуляризаторы — дополнительные критерии $R_i(\Phi, \Theta)$, характеризующие качество модели. Например, регуляризаторы могут задавать лингвистические свойства, которыми должна обладать модель.

Для оценивания параметров модели максимизируется логарифм правдоподобия со взвешенной суммой регуляризаторов $R(\Phi, \Theta) = \sum\limits_i \tau_i R_i(\Phi, \Theta)$:

$$\sum_{m\in M} \eta_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} ln \sum_{s\in S} \varphi_{ws}\theta_{sd} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}, \tag{1}$$

$$\sum_{w\in W^m} \varphi_{ws} = 1, \varphi_{ws} \geqslant 0; \ \sum_s \theta_{sd} = 1, \theta_{sd} \geqslant 0, \tag{2}$$

где коэффициенты η_m введены для балансирования важности модальностей W^m, а коэффициенты τ_i – для балансирования между оптимизацией нескольких критериев.

3 Иерархическая тематическая модель

3.1 Обозначения и определения

При фиксированном уровне l иерархии будем обозначать S – множество тем данного уровня, T – множество родительских тем, то есть множеством тем $(l-1)$-го уровня. Матрицы родительского уровня будем обозначать $\Phi^p \in R^{|W|\times|T|}$ и $\Theta^d = R^{|T|\times|D|}$.

Обозначим $\Psi = \{\psi_{st}\}_{s\in S, t\in T}$, $\psi_{st} = p(s|t)$, и будем называть ее матрицей перехода между уровнями. Величина ψ_{st} показывает вероятность перехода в подтему S из надтемы t.

3.2 Послойное построение иерархии

Данная модель основана приближением матрицы Φ^p родительского уровня произведением $\Phi\Theta$:

$$p(w|t) \approx \sum_s p(w|s)p(s|t) = \sum_s \varphi_{ws}\psi_{st}. \tag{3}$$

Данное разложение подразумевает введение еще одной гипотезы условной независимости $p(w|s,t) = p(w|s)$: вероятность появления термина w в подтеме s не зависит от надтемы t.

Оптимизационная задача имеет вид:

$$lnL(\Phi, \Theta, \Psi) = \sum_{n_d w} ln \sum_s \varphi_{ws}\theta_{sd} + \lambda \sum_t \sum_w \varphi^p_{wt} ln \sum_{s\in S} \varphi_{ws}\psi_{st} + R(\Phi, \Theta, \Psi) \to \max_{\Phi,\Theta,\Psi}, \tag{4}$$

$$\sum_{w\in W^m} \varphi_{ws} = 1, \varphi_{ws} \geqslant 0; \ \sum_s \theta_{sd} = 1, \theta_{sd} \geqslant 0; \ \sum_s \psi_{st} = 1, \psi_{st} \geqslant 0. \tag{5}$$

При таком матричном разложении $\Phi^p = \Phi\Psi$ мы требуем, чтобы столбцы родительской Φ^P были линейной комбинацией столбцов дочерней Φ, то есть представляем родительскую тему как смесь дочерних тем.

Формула описанного регуляризатора имеет схожую структуру с формулой правдоподобия модели. Поэтому такая постановка задачи эквивалентна добавлению во входную матрицу $n_{dw}|T|$ псведодокументов, отвечающих столбцам родительской Φ: $n_{d'w} = \lambda\varphi_{wt}, d' = t$, и для обучения модели применим стандартный EM-алгоритм. Отвечающая введенным псевдодокументам часть Θ составит матрицу Φ.

4 Функционалы качества первого уровня

Пусть Φ^g – подматрица матрицы Φ, соответствующая модальности тегов, Φ^w – подматрица матрицы Φ, соответствующая остальным модальностям.

Пусть S_s – множество предметных тем, S_b – множество фоновых тем, $S = S_s \cup S_b$, G – множество тегов, G_d – множество тегов документа d. Приэтом $S_s \subset G$. В общем случае каждая тема может быть представлена несколькими тегами.

Фоновость документа:

$$b(d) = \sum_{s \in S_b} \theta_{sd}.$$

Для первого уровня мы можем сами задать какой тег соответствует каждой теме (предполагаем, что каждая тема представлена единственным тегом). Для этого соответствующую подматрицу матрицы Φ^g инициализируем единичной до построения тематической модели. В результате элементы матрицы Φ^g, соответсвующие столбцам φ_s, $s \in S_s$: 0 или 1.

4.1 Ошибка второго рода

FNR – доля ошибок вида: тег был в документе d ($g_d = s \in G_d \cap S_s$), а его темы (темы, соответствующей данному тегу) нет: $\theta_{sd} = p(s|d) = 0$.

$$FNR = \frac{\sum\limits_d \sum\limits_{s \in G_d \cap S_s} [\frac{\theta_{sd}}{1-b(d)} < k]}{\sum\limits_d |G_d \cap S_s|},$$

где k – заданный порог.

4.2 Ошибка первого рода

FPR – доля ошибок вида: тега нет в документе d ($g_d = s \in S_s \setminus G_d$), но тема есть $\theta_{sd} = p(s|d) = 1$.

$$FPR = \frac{\sum\limits_d \sum\limits_{s \in S_s \setminus G_d} [\frac{\theta_{sd}}{1-b(d)} \geqslant k]}{\sum\limits_d |S_s \setminus G_d|},$$

где k – заданный порог.

4.3 Функционал

Смысл функционала: доля тем, которые действительно есть в документах.

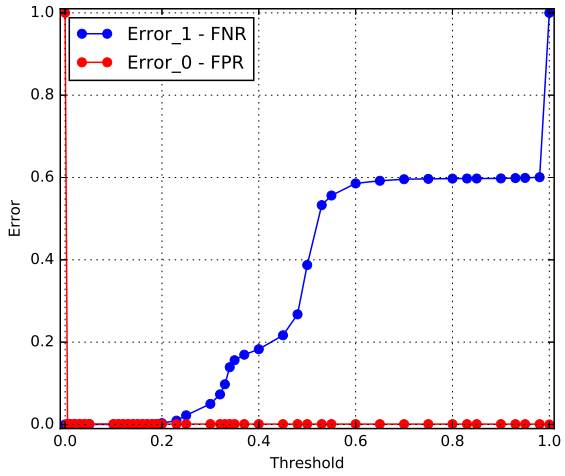
$$B = \frac{1}{\sum\limits_d |G_d \cap S_s|} \sum_d \sum_{g \in G_d} \sum_{s \in S_s : p(g|S) \neq 0} \frac{\theta_{sd}}{1 - b(d)},$$

Функционал B при идеальном построении модели равен 1.

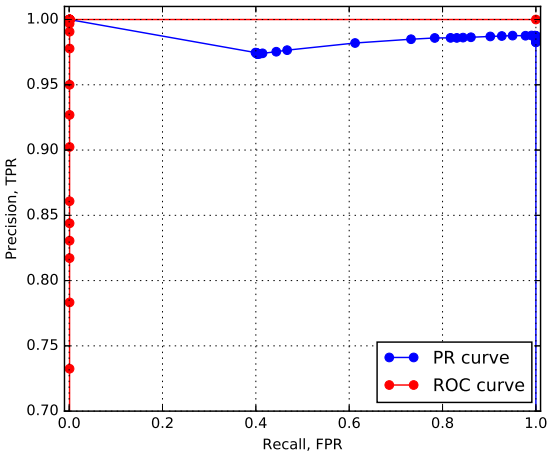
5 Эксперименты

Для проведения экспериментов подготовлена лемматизированная и преобразованная в матрицу частот слов коллекция научно-популярных текстов с сайта `https://postnauka.ru/`. В коллекции 3446 документов, размер словаря моадльности слов – 19187, модальности биграмм – 11633, трехграмм – 740, четыреграмм – 85, авторов – 860, тегов – 930. Энграммы выделены с использованием внешних средств.

Результаты экспериментов представлены на графиках.

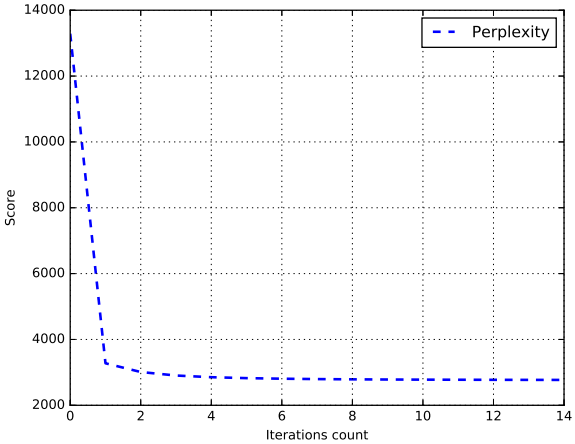


(*а*) График зависимости FNR и FPR от значения порога.

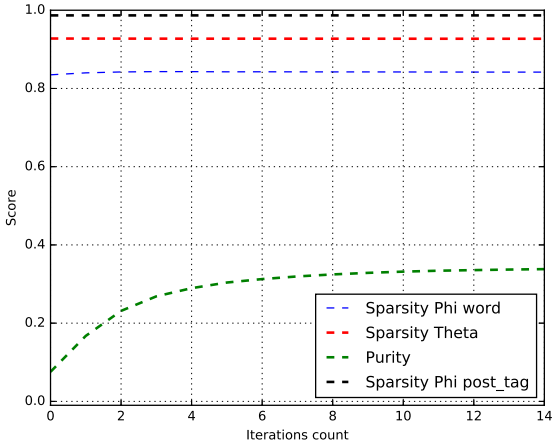


(*б*) График Precision-Recall кривой и ROC-кривой.

Рис. 1



(*а*) График зависимости перплексии от номера итерации алгоритма.



(*б*) Графики зависимости разреженности матриц от номера итерации алгоритма.

Рис. 2

6 Заключение

Предложен метод послойного построения иерархии на коллекциях, для которых имеются экспертные знания о наличии тем в документах и дополнительные модальности, для решения задачи построения иерархической мультимодальной тематической модели

коллекции научно-популярных текстов. Также предложены метрики качества для оценки отдельных уровней иерархии. Метод реализован для первого уровня иерархии. Планируется:

– реализовать метод для второго и третьего уровней иерархии,
– разработать и применить методы оценки качества построенной иерархической тематической модели,
– предложить и реализовать методы автоматического определения числа тем и их именования; выбора новой темы при появлении документа в коллекции, не похожего на уже имеющиеся.

Литература

[1] *Blei M., Griffiths T., Jordan M., Tenenbaum J.* Hierarchical topic models and the nested Chinese restaurant process // NIPS, 2003.

[2] *David M., Li W., McCallum A.* Mixtures of Hierarchical Topics with Pachinko Allocation // ICML, 2007.

[3] *Zavitsanos E., Paliouras G., Vouros G.* Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // Mach. Learn. Res., Novemver 2011. Vol. 12, P. 2749–2775.

[4] *Wang C., Danilevsky M., Desai N.* A Phrase Mining Framework for Recursive Construction of a Topical Hierarchy // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, – New York, USA : ACM, 2013. Vol. 12, P. 437–445.

[5] *Wang C., Liu J., Desai N.* Constructing topical hierarchies in heterogeneous information networks // Knowledge and Information Systems, 2014. Vol. 44, No. 3, P. 529–558.

[6] *Wang C., Liu X., Song Y.* Scalable and Robust Construction of Topical Hierarchies // CoRR, 2014. Vol. abs/1403.3460.

[7] *Wang C., Liu X., Song Y., Han J.* Towards Interactive Construction of Topical Hierarchy: A Recursive Tensor Decomposition Approach // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, – New York, USA : ACM, 2015. P. 1225–1234.

[8] *Pujara J., Skomoroch P.* Towards Interactive Construction of Topical Hierarchy: A Recursive Tensor Decomposition Approach // NIPS Workshop on Big Learning, 2012.

[9] *Chirkova N., Vorontsov K.* Additive Regularization for Hierarchical Multimodal Topic Modeling // Machine Learning and Data Analysis,, 2016. Vol. 2, Issue. 1.

[10] *Vorontsov K., Potapenko A.* Analysis of Images // Social Networks and Texts: Third International Conference, AIST, 2014, Yekaterinburg, Russia, April 10-12, 2014, Revised Selected 39 Papers / Под ред. I. Dmitry Ignatov, Yu. Mikhail Khachay, Alexander Panchenko и др. — Cham : Springer International Publishing, 2014. — С. 29–46. — ISBN: 978-3-319-12580-0.